ГЛАВА XX.
Поиск аномалий

Я против гомогенизированного общества,
потому что я хочу, чтобы сливки всплывали.
Р. Фрост

В других нас раздражает не отсутствие
совершенства, а отсутствие сходства с нами.
Д. Сантаяна

Поговорим об одной важной проблеме обучения на неразмеченных данных (Unsupervised Learning) – задаче **поиска аномалий (Anomaly Detection)**. Строго говоря, в анализе данных есть два направления, которые занимаются поиском аномалий: **детектирование выбросов (Outlier Detection)** и **«новизны» (Novelty Detection)**. И **выброс и новизна – это объект, который отличается по своим свойствам от объектов (обучающей) выборки**. Но в задаче обнаружения выбросов предполагается, что эти выбросы есть в выборке, а в задаче обнаружения новизны нам дана выборка без выбросов – однородных объектов – и требуется разработать алгоритм, который по описанию нового объекта определит, насколько он похож на объекты выборки (не является ли он «новым объектом», см. рис. XX.1). Например, если вы анализируете замеры температуры и отбрасываете аномально большие и маленькие, то вы боретесь с выбросами. А если вы создаёте алгоритм, который для каждого нового замера оценивает, насколько он похож на прошлые, и выбрасывает аномальные, то «боретесь с новизной».

Выбросы уже есть в выборке, новизна – скоро появится.

Аномалии являются следствием:

- ошибок в данных (неточности измерения, округления, неверной записи и т.п.),
- наличия шумовых объектов (неверно классифицированных объектов),
- присутствия объектов «других» выборок (например, датчик ломается и в выборку добавляются показания сломавшегося датчика).

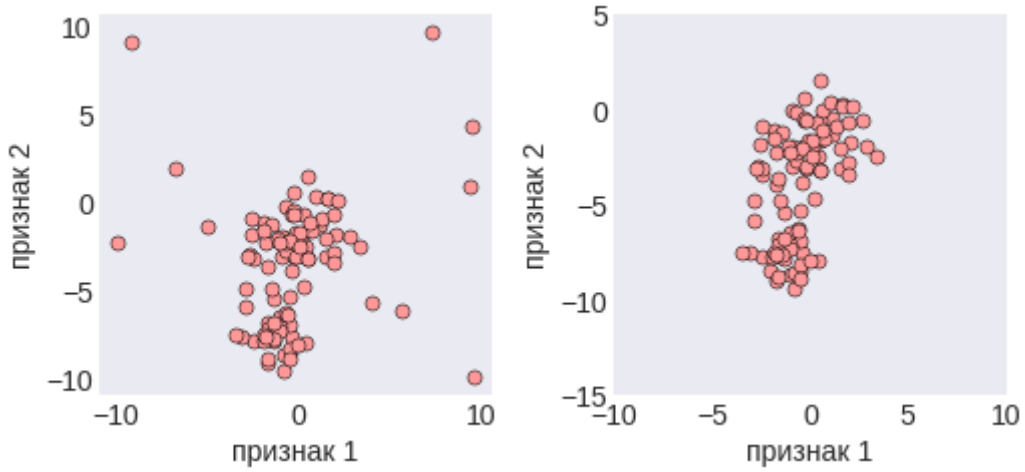

Рис. XX.1. Обучающая выборка в задаче обнаружения выбросов (слева) и новизны (справа).

Вообще, выбросы могут быть таковыми «в слабом смысле», когда ошибки в описании объектов незначительны, и объекты просто оказываются на границах кластеров точек, см. рис. XX.2 (также говорят, что «шум размывает границы»), а могут быть «в сильном смысле», когда они в признаковом пространстве сильно не похожи на остальные объекты выборки. Нас прежде всего интересуют такие аномалии, т.к. они влияют на остальные алгоритмы машинного обучения (например, потому что искажают границы классов в задаче с размеченными данными[1]).

Шум (noise) – это выброс «в слабом смысле».

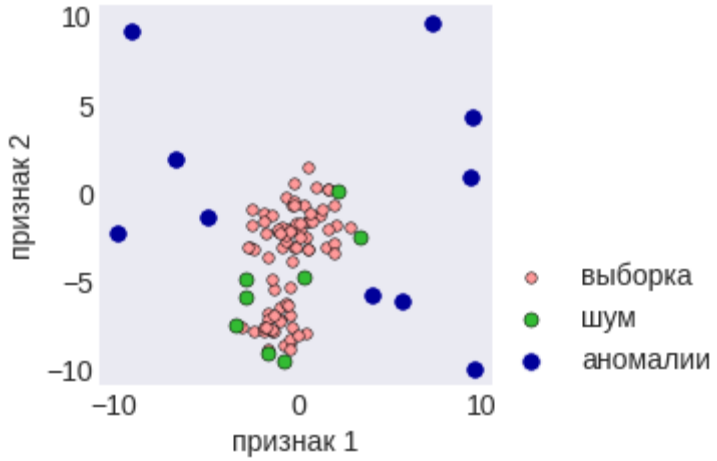

Рис. XX.2. Модельная задача обнаружения аномалий с двумя признаками.

Новизна (новые объекты), как правило, появляется в результате принципиально нового поведения системы / замера или сбора признаков. Скажем, если наши

[1] Обратим внимание, что обнаружение аномалий, как и другие задачи USL, могут решаться на данных без разметки, но это не значит, что такой разметки нет.

объекты – описания работы системы, то после проникновения в неё вируса объекты становятся «новизной». Ещё пример – описания работы двигателя после поломки. Обратим внимание на термин **«новизна»** – похожих объектов в достаточном количестве не было в обучающей выборке. Также обратим внимание, что их часто и нельзя «наколлекционировать», например получить описания работы двигателя при всевозможных поломках. Формирование такой обучающей выборки трудозатратно и часто не имеет смысла. На рис. XX.2, например, нет выброса, который лежит «внизу выборки». Зато можно набрать достаточно большую выборку примеров нормальной (штатной) работы системы или механизма.

У поиска аномалий довольно много приложений, кроме того, что часто очистка от выбросов – естественный этап предобработки данных:

- финансовая аналитика, обнаружение подозрительных банковских операций (Credit-card Fraud),
- обнаружение нестандартных игроков на бирже (инсайдеров), аномалии торгового поведения (trading anomaly),
- кибербезопасность, анализ аномальности сетевого трафика, обнаружение вторжений (Intrusion Detection),
- мониторинг оборудования и промышленных систем, обнаружение неполадок в механизмах по показаниям датчиков,
- медицинская диагностика (Medical Diagnosis),
- сейсмология.

Стоит отметить, что есть несколько возможных постановок задач обнаружения выбросов. На рис. XX.3 (слева) показана задача в типичной постановке обучения на неразмеченных данных: дана выборка без меток, необходимо найти все выбросы. На рис. XX.3 (центр) показана часто встречающаяся постановка: **Positive-Unlabeled Classification (PU learning)** – когда часть выбросов размечена (класс 1), но в остальных объектах обучения (класс 0) также могут содержаться выбросы. Например, нам эксперт сказал, что оборудование давало сбой в такие-то моменты времени, но он мог заметить не все сбои. Также возможна задача с полной разметкой (рис. XX.3), когда все аномалии помечены меткой «1», а остальные объекты – меткой «0», тогда это обычная задача классификации. Однако, такая постановка редкость (вряд ли Вы гарантировано нашли в системе все вирусы, поломки и т.п.). В задаче

обнаружения новизны только одна возможная постановка: по неразмеченной выборке обучить алгоритм детектирования новых объектов, поскольку здесь в выборке посторонних объектов нет.

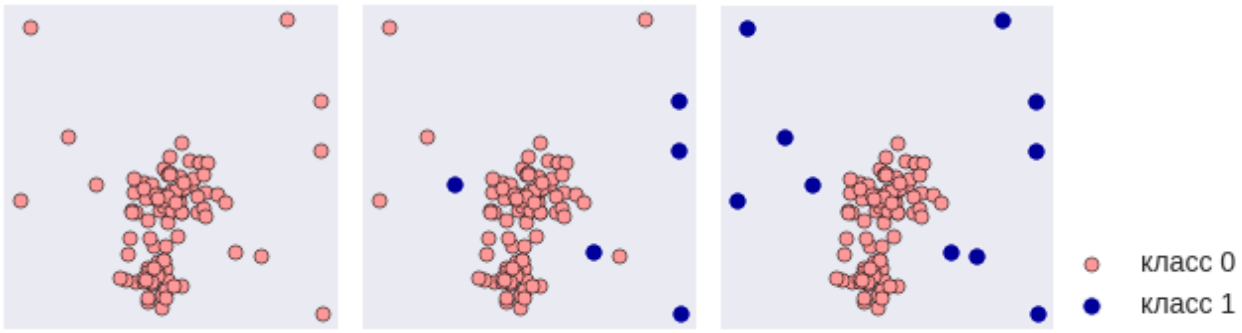

Рис. XX.3. Разные постановки задачи: без разметки, с частичной разметкой класса 1, с полной разметкой.

В задачах обнаружения аномалий, как правило, есть дисбаланс классов (например, поломки оборудования и мошеннические транзакции относительно редки). Аномалии бывают не только в табличных данных, они могут быть в графах (выбросами могут быть вершины, рёбра и сами графы), временных рядах, строках и т.д. (см. рис. XX.4-6).

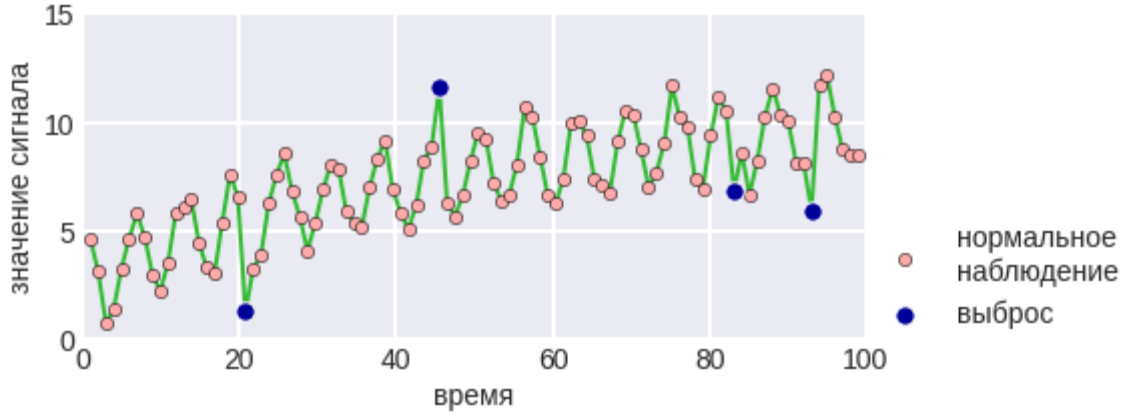

Рис. XX.4. Временной ряд и выбросы в нём.

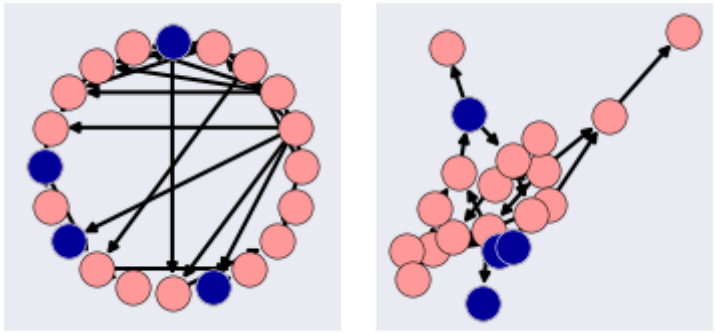

Рис. XX.5. Выбросы-вершины на графах.

AAABBCCAABBBCAAABBCABBCCABAABBCCAAABBBBCABBBCC
BBCCAACCBBCCAABBABAABBBBCCAABBBCBBCCACAABB

Рис. XX.6. Пример выбросов в последовательностях.

Функционалы качества в задачах детектирования аномалий используют примерно такие же, как и в задачах классификации: PR AUC, AUROC, здесь всё определяется контекстом задачи (заказчиком).

Далее рассмотрим основные группы методов обнаружения аномалий.

Статистические тесты

Статтесты, как правило, применяют для отдельных признаков и отлавливают экстремальные значения (Extreme-Value Analysis). Для этого используют, например, Z-value:

$$Z_i = \frac{|x_i - \mu|}{\sigma},$$

μ – выборочное среднее, σ – выборочное стандартное отклонение (сравнивают с порогом), а по значению Kurtosis measure

$$\frac{1}{m}\sum_{i=1}^{m} Z_i^4$$

определяют наличие выбросов в выборке. Есть простые методы, например оценивают плотность с помощью непараметрического метода (гистограммным методом или по Парзену), точки, которые лежат в области небольших значений плотности объявляются выбросами, см. рис. XX.7. Здесь и в методах, о которых расскажем дальше, задаются гиперпараметром «доля ожидаемых выбросов» и исходя из него подбирают другие параметры, например, какое значение оценки плотности считать небольшим.

Любой практик имеет какой-нибудь свой проверенный способ нахождения экстремальных значений для определённых типов данных. Многие методы визуализации, например ящик с усами (box-plot), имеют встроенные средства для детектирования и показа таких экстремальных значений, см. рис. XX.8. Напомним, что ящик заполняет пространство от первой Q1 до третей квартили Q3, середина ящика – медиана Q2, а усы отмечают Q2 ± 1.5IRQ, где IRQ = Q2 –

Q1 – интерквартильный размах. Точки, которые выходят за пределы усов считаются выбросами.

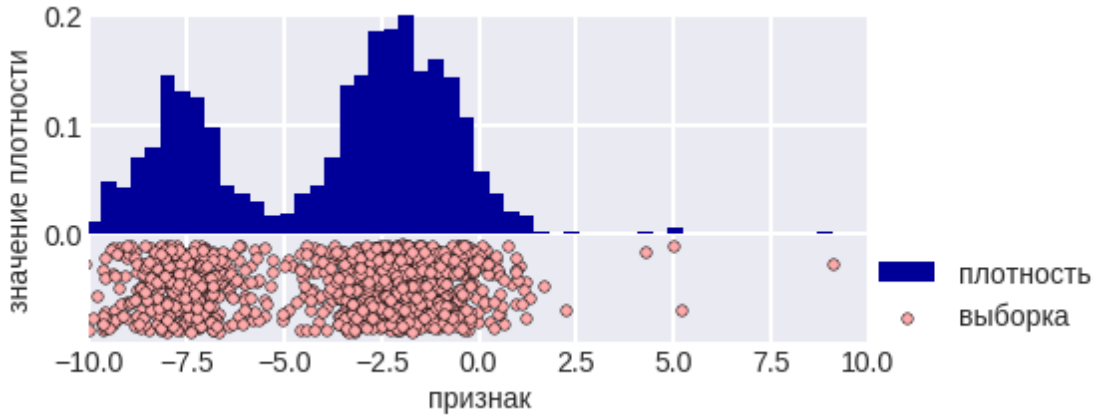

Рис. XX.7. Оценка плотности.

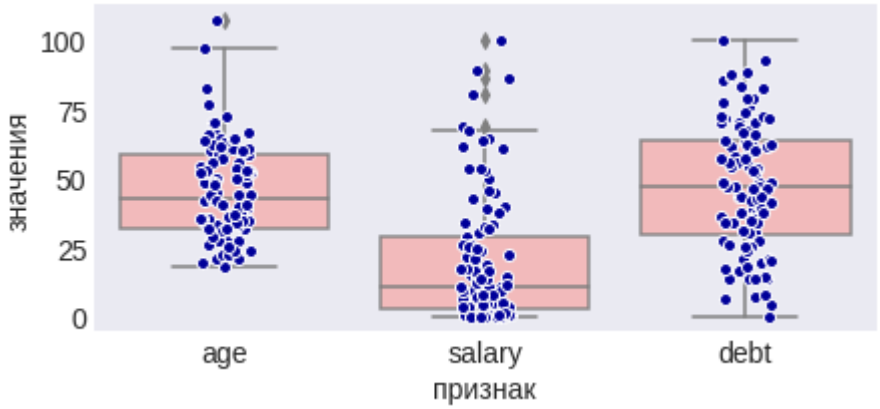

Рис. XX.8. Примеры ящиков с усами для отдельных признаков.

Важно понимать, что экстремальное значение и аномалия это разные понятия. Например, в выборках

```
[10,  2,  0, 16,  6, 20,  4,  6, 10, 16, 10,  1,  6, 10, 14, 16,  2, 0, 10, 16]
[31,  4, 33,  5, 33,  4, 30,  5, 30, 21,  5,  0,  4, 34, 32, 31, 34, 32, 31, 35]
```

отмечены аномальные значения. В первой такое значение отличается от остальных чётностью, во второй – тем, что в её достаточно большую окрестность не попадают другие точки выборки. При этом такие значения не являются минимальными или максимальными. На рис. XX.9 показана задача с двумя признаками, визуально три объекта отличаются от остальных объектов выборки, но при этом по каждому из признаков они лежат внутри достаточно плотно заполненного точками отрезка. По отдельным признакам эти объекты не отличаются от остальных объектов выборки.

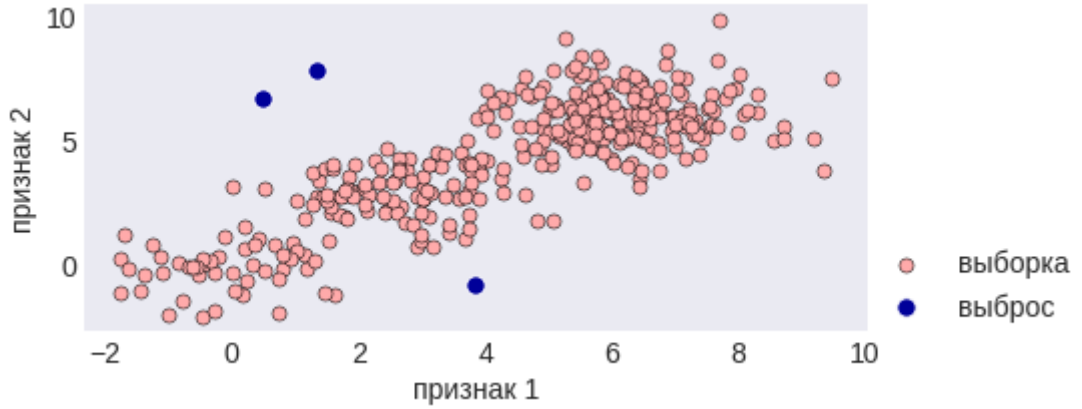

Рис. XX.9. Пример выбросов в задаче с двумя признаками.

Модельные тесты

Идея очень простая – мы строим модель, которая описывает данные. Точки, которые сильно отклоняются от ответов модели (на которых модель сильно ошибается) и есть аномалии (см. рис. XX.10). При выборе и обучении модели мы можем учесть природу задачи, функционал качества и т.п. Такие методы хороши для определения новизны, но хуже работают при поиске выбросов, поскольку модель строится не по идеальной выборке, а выборке, испорченной выбросами, которые мы и должны найти.

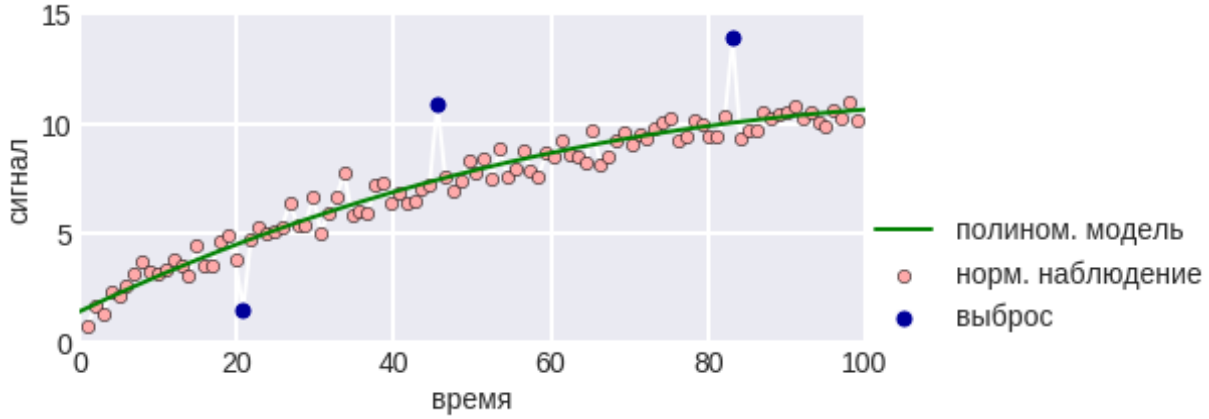

Рис. XX.10. Описание данных моделью и нахождение точек, которые от неё максимально отклоняются.

Модельный подход очень универсальный и подходит для задач с объектами разной природы. На рис. XX.11 показано применение модельного подхода для поиска выбросов в элементах матрицы. Это не обязательно матрица данных (объект-признак). Мы используем неполное сингулярное разложение (SVD), чтобы найти матрицу небольшого ранга максимально похожую на нашу:

$$X = \| x_{ij} \|_{m \times n} \approx H_k = \| h_{ij} \|_{m \times n} = U_{m \times k} L_{k \times k} V_{k \times n},$$

для каждого ij-элемента посчитаем ошибку приближения матрицы с помощью сингулярного разложения:

$$e_{ij} = (x_{ij} - h_{ij})^2.$$

Элементы, которые сильно отличаются от соответствующих элементов матрицы небольшого ранга, будем считать выбросами. На рис. XX.11 интенсивностью показана ошибка e_{ij} при $k = 2$ (это, кстати, гиперпараметр метода). Элемент «6» максимально интенсивно подсвечен (видимо потому, что в его строке остальные элементы существенно меньше).

	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
x_0	0	5	6	4	6	1	10	6	5	1
x_1	5	7	10	6	7	5	9	9	2	5
x_2	2	4	5	5	4	4	8	5	9	4
x_3	3	5	5	4	8	3	7	5	9	1
x_4	0	6	2	2	1	3	3	7	0	2
x_5	1	8	14	12	7	6	12	6	11	7

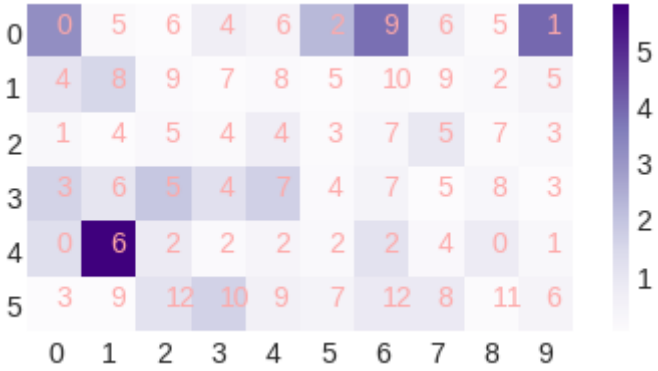

Рис. XX.11. Применение SVD для нахождение выбросов в матрице

Итерационные методы

Есть много эвристических итерационных методов, на каждой итерации которых удаляется группа «особо подозрительных объектов». Например, в n-мерном признаковом пространстве можно удалять выпуклую оболочку наших точек-объектов, считая её представителей выбросами. В одномерном случае, если постоянно отбрасывать максимальный и минимальный элемент, то в итоге в выборке останется медиана (или пара точек, среднее которых является

медианой). В многомерном – последовательное удаление точек на выпуклой оболочке позволяет ввести понятие многомерной медианы, а сами выпуклые оболочки описывают экстремальные точки, первая – выбросы первого порядка, вторая – второго и т.д., см. рис. XX.12. Как правило, методы этой группы достаточно трудоёмки.

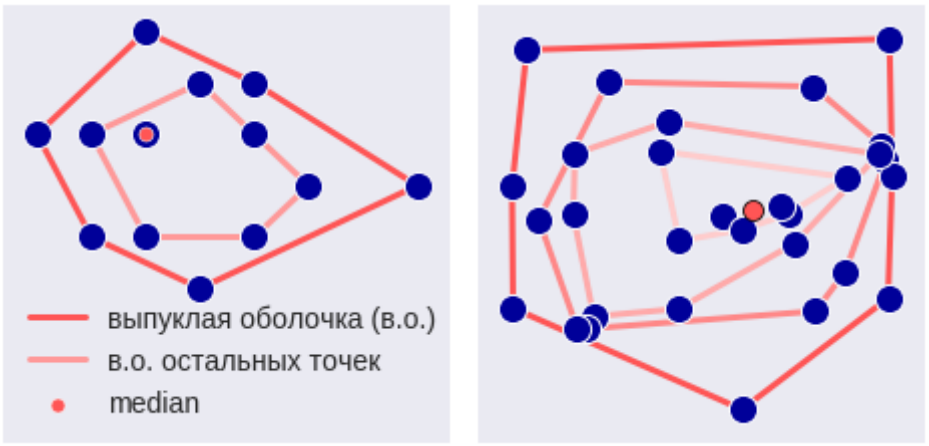

Рис. XX.12. Выпуклые оболочки множества точек.

Метрические методы

Судя по числу публикаций, методы в которых анализируются расстояния между объектами самые популярные среди исследователей. В них постулируется существование некоторой метрики в пространстве объектов, которая и помогает найти аномалии. Интуитивно понятно, что у выброса мало соседей в окрестности фиксированного радиуса, а у типичной точки много. Ну или от выброса соседи (ближайшие точки выборки) больше удалены, см. рис. XX.13. Поэтому хорошей мерой аномальности может служить, например «расстояние до k-го соседа». Часто используются специфические метрики, например расстояние Махаланобиса.

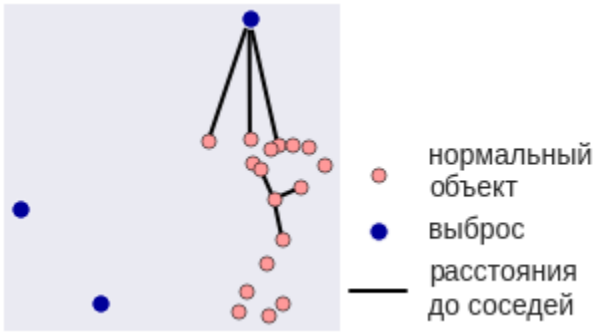

Рис. XX.13. Соседи разных элементов выборки.

Опишем метод **Local Outlier Factor (LOF[1])**. Пусть $\rho_k(z)$ – расстояние до k-го ближайшего соседа из выборки, будем использовать «поправленное[2]» расстояние

$$\rho'(x,z) = \max[\rho(x,z), \rho_k(z)].$$

Для точки x посчитаем усреднение поправленных расстояний до соседей:

$$r_k(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} \rho'(x, x_i),$$

обратная величина к этой называется **local reachability density (LRD)**. Тогда оценка «выбросовости» **relative distance score** вводится как:

$$\frac{1}{k} \sum_{x_i \in N_k(x)} \frac{r_k(x)}{r_k(x_i)}.$$

Заметим, что она тем больше, чем дальше точка от своего k-го ближайшего соседа из выборки, при этом мы нормируем на такие подобные расстояния до соседей. На рис. XX.14 показаны оценки аномальности для точечной конфигурации на плоскости.

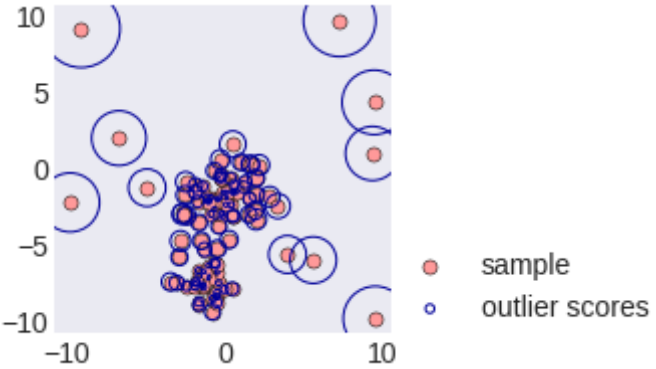

Рис. XX.14. Выборка и оценки выбросовости для точек (соответствую радиусам окружностей с центрами в точках).

Методы подмены задачи

Когда возникает новая задача, есть большой соблазн решить её старыми методами (ориентированными на уже известные задачи). Например, можно

[1] Метод предложен в Breunig M. M. et al. LOF: identifying density-based local outliers //Proceedings of the 2000 ACM SIGMOD international conference on Management of data. – 2000. – С. 93-104.

[2] Иногда используют обычное расстояние.

сделать кластеризацию, тогда маленькие кластеры, скорее всего, состоят из аномалий, см. рис. XX.15. Также можно сделать кластеризацию и расстояние точки до ближайшего центра кластера считать оценкой аномальности, см. рис. XX.16 (тут как раз можно использовать расстояние Махаланобиса). Некоторые кластеризаторы имеют встроенную технику детектирования выбросов, например DBSCAN (см. главу про кластеризацию).

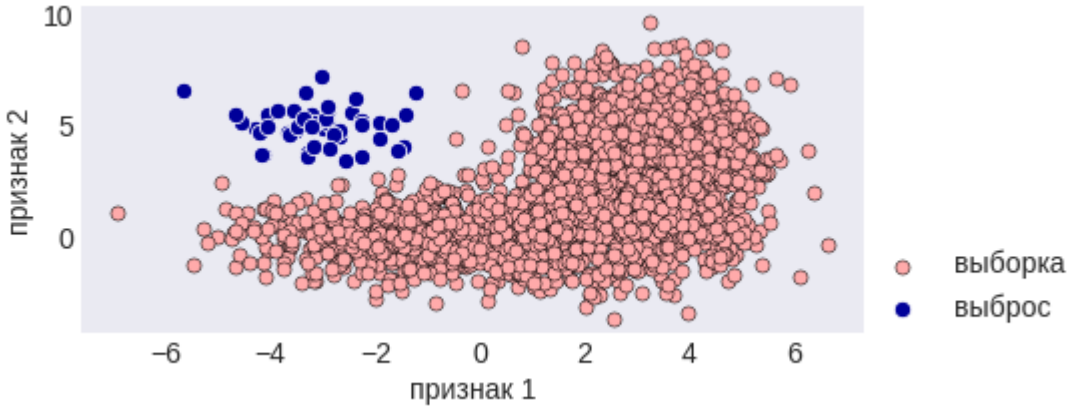

Рис. XX.15. Пример кластеризации на малый (красный) и большой (синий) кластер.

Если у нас есть частичная информация об аномалиях (как в постановке PU learning), то можно решить её как задачу классификации с классами 1 (размеченные аномалии) и 0 (все остальные объекты). Если бы класс 0 состоял только из нормальных объектов, то такое решение было бы совсем законным, иначе остаётся надеяться, что недетектированных аномалий в нём немного.

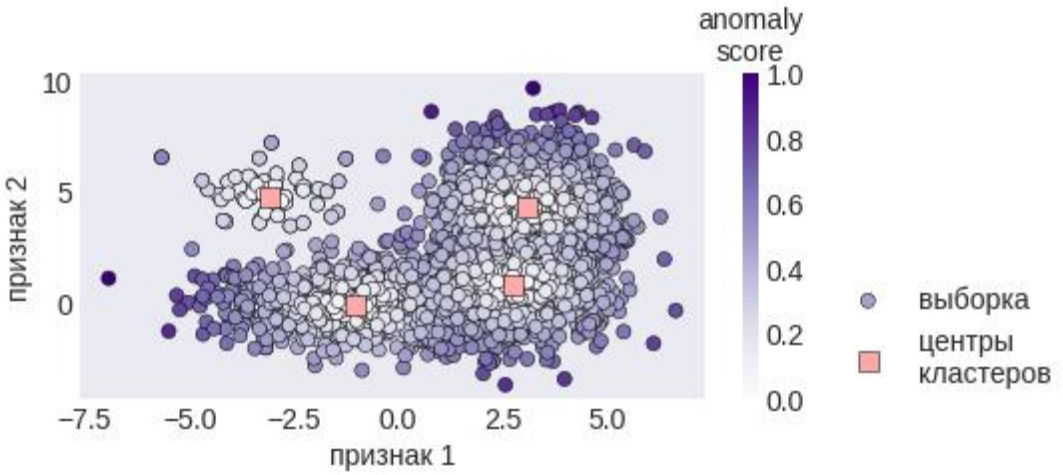

Рис. XX.16. Оценка аномальности как нормированное расстояние до ближайшего центра кластера.

Методы машинного обучения

А что если воспринять задачу нахождения аномалий как новую задачу машинного обучения (отличную от классификации и кластеризации)? Самые популярные алгоритмы[1] здесь:

- метод опорных векторов для одного класса (OneClassSVM[2]),
- изолирующий лес (IsolationForest[3]),
- эллипсоидальная аппроксимация данных (EllipticEnvelope[4]).

Все методы имеют гиперпараметры, которые определяют, какую долю объектов относить к выбросам.

Первый метод — это обычный SVM, который отделяет выборку от начала координат. Идея немного сомнительна, но оказалась довольно работоспособной (см. рис. XX.16). Здесь правда не так много разнообразия в выборе параметров, как при решении задач классификации, поскольку на практике в качестве ядра подходит лишь rbf (радиальные базисные функции), все остальные ядра показывают феноменально плохой результат, см. рис. XX.17. Интересно , что многие годы задачи детектирования поломок сложных механизмов решались именно с помощью OneClassSVM, почему-то без рассмотрения альтернатив. Полезно помнить, что **OneClassSVM это скорее алгоритм поиска новизны**, а не выбросов, т.к. «затачивается» под обучающую выборку.

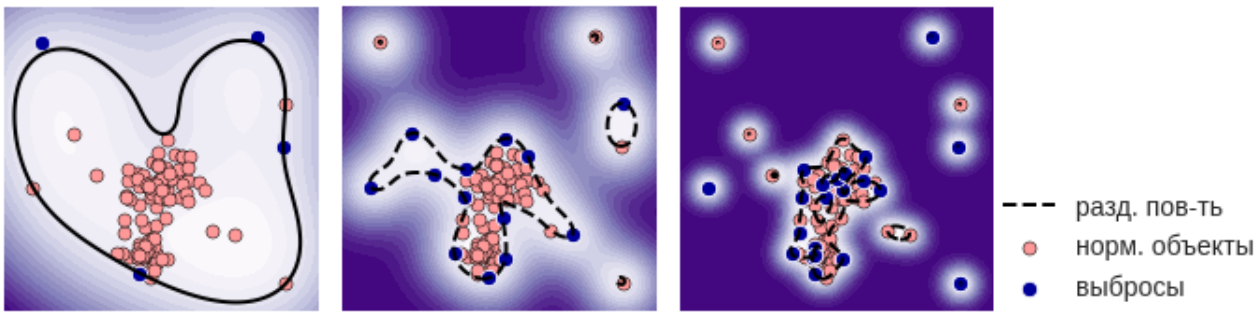

Рис. XX.16. Результат работы метода OneClassSVM с ядром rbf при разных значениях гиперпараметра gamma = 0.01, 0.1, 0.5[1].

[1] Есть реализация в scikit-learn.

[2] Schölkopf B. et al. Support vector method for novelty detection //Advances in neural information processing systems. – 1999. – Т. 12.

[3] Liu F. T., Ting K. M., Zhou Z. H. Isolation forest //2008 eighth ieee international conference on data mining. – IEEE, 2008. – С. 413-422.

[4] EStimator D. A Fast Algorithm for the Minimum Covariance //Technometrics. – 1999. – Т. 41. – №. 3. – С. 212.

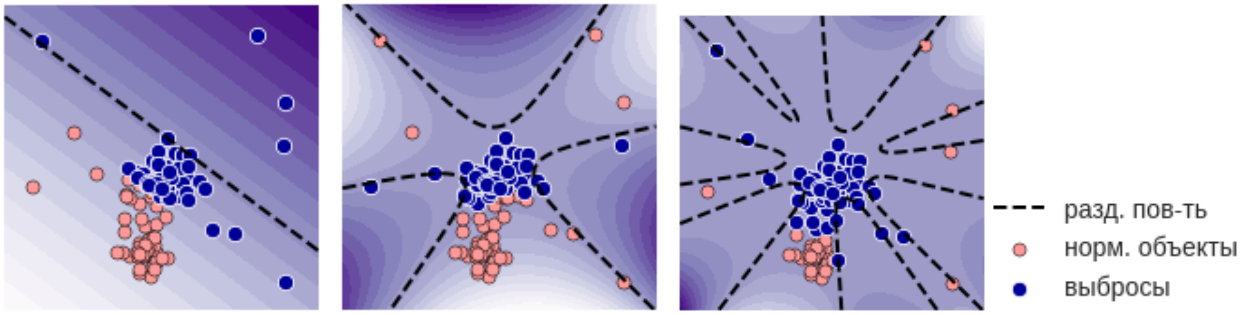

Рис. XX.17. Результат работы метода OneClassSVM с ядром poly при разных значениях гиперпараметра degree = 1, 5, 9.

Изолирующий лес[2] **(Isolation Forest)** – это одна из вариаций идеи случайного леса, как всегда, простая и надёжная:

- лес состоит из отдельных базовых алгоритмов – деревьев, их число – гиперпараметр метода n_estimators,
- каждое дерево строится до исчерпании выборки (пока ровно один объект не останется в листе), при построении дерева можно использовать часть выборки, её объём определяется гиперпараметром max_samples, можно использовать бутстреп и/или случайное подмножество признаков (число определяется гиперпараметром max_features),
- для построения ветвления в дереве выбирается случайный признак и случайное расщепление по нему (случайный порог),
- для каждого объекта мера его нормальности – среднее арифметическое глубин листьев, в которые он попал (изолировался), порог, ниже которого объект считается выбросом, определяется из значения гиперпараметра contamination – ожидаемая доля выбросов в выборке.

Идея алгоритма простая: при описанном случайном способе построения деревьев выбросы будут попадать в листья на ранних этапах (на небольшой глубине дерева), т.е. выбросы проще «изолировать» (напомним, что дерево строится до тех пор, пока каждый объект не окажется в отдельном листе), см. рис. XX.18. Алгоритм хорошо отлавливает именно выбросы (см. рис. XX.19-21).

[1] На этом рисунке и далее, цветом обозначены предсказания модели, гиперпараметр «доля выбросов» которой равнялся 0.05. Разделяющая поверхность (она пунктирная, но на некоторых рисунках сливается в сплошную) соответствует 5-персентилю оценки нормальности объектов обучения.

[2] Более корректный перевод термина «Изоляционный лес», но используемый нами передаёт принцип построения леса: в деревьях изолируются точки.

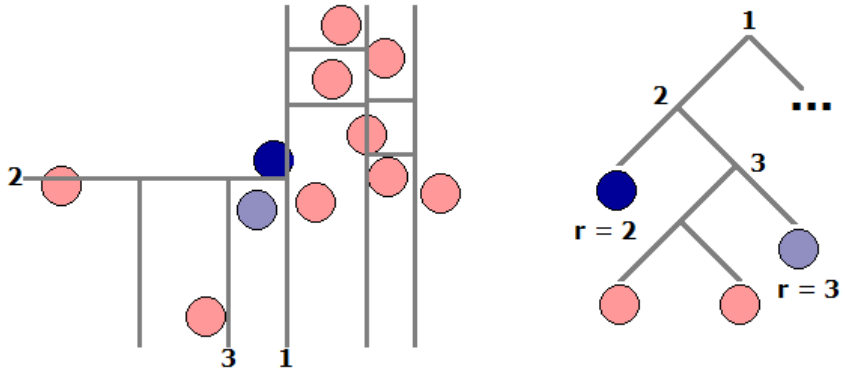

Рис. XX.18. Вычисление оценки аномальности в изолирующем лесе (синий объект изолировался на втором уровне, светлосиний – на третьем).

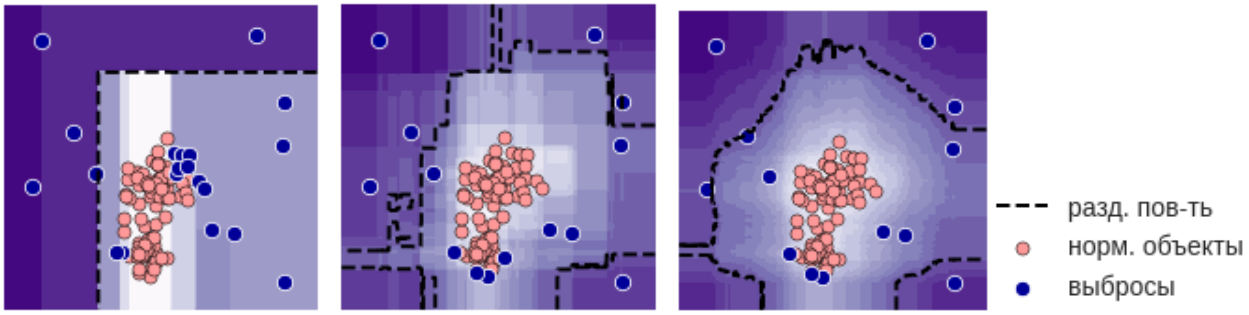

Рис. XX.19. Оценка аномальности, полученная изолирующим лесом при числе деревьев = 1, 10, 100.

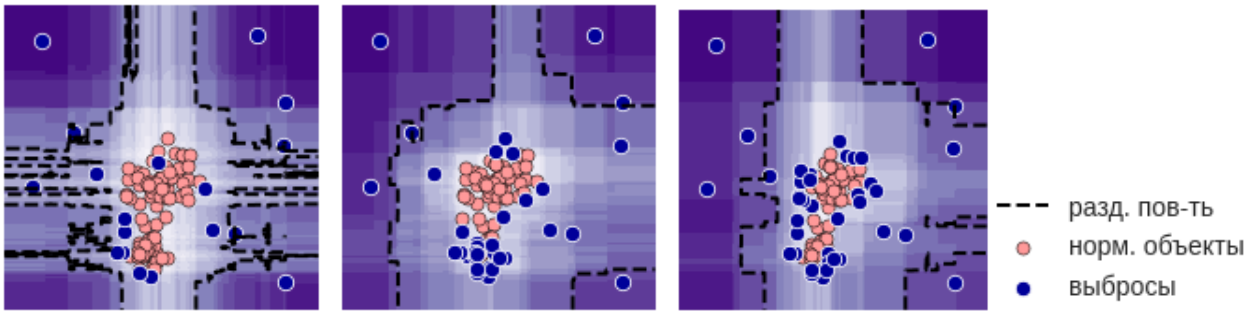

Рис. XX.20. Оценка аномальности, полученная изолирующим лесом при max_features=0.5 (слева), bootstrap=True (по центру), max_samples=0.5 (справа).

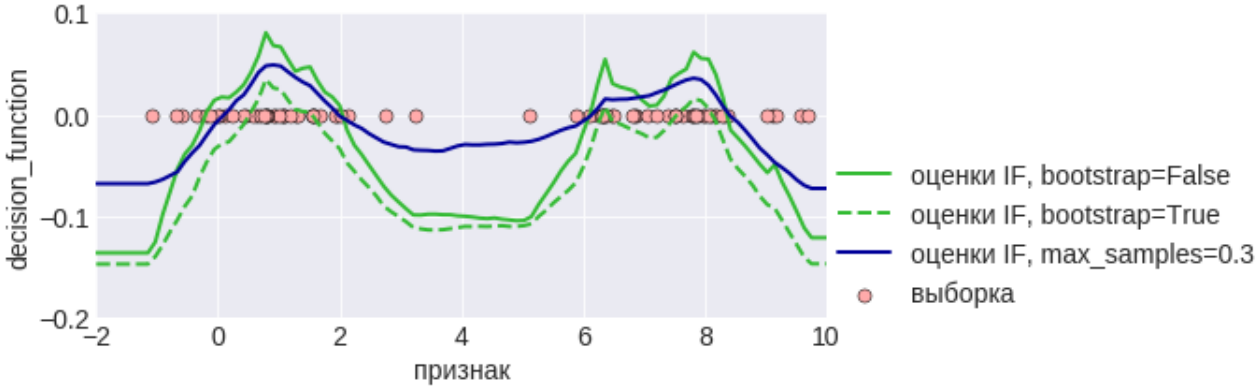

Рис. XX.21. Оценки аномальности, полученные изолирующим лесом по одномерной выборке.

В методе **эллипсоидальной аппроксимации данных** строится эллипсоид, который пытается «захватить» облако точек, см. рис. XX.22. Метод хорошо работает только на одномодальных данных, а совсем хорошо – на нормально распределённых. Степень новизны здесь фактически определяется по расстоянию Махалонобиса.

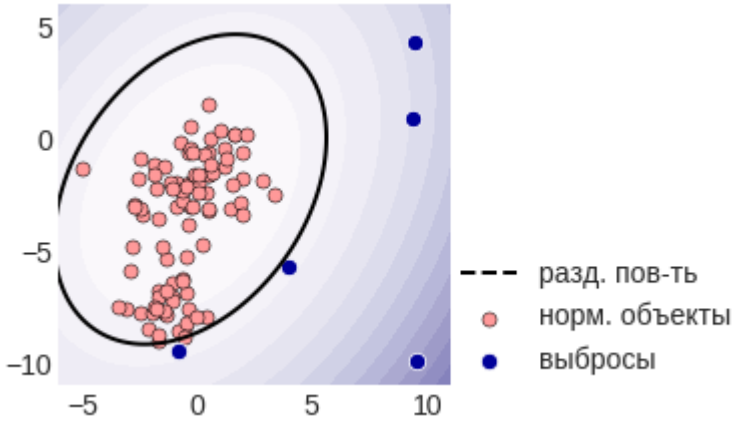

Рис. XX.22. Построение эллипса, описывающего облако точек.

Ансамбли алгоритмов

В область обнаружения аномалий также проникла идея «один алгоритм хорошо, а сто лучше», поэтому часто строят много разных алгоритмов. Каждый из них даёт оценку аномальности и эти оценки потом «усредняют».

Поскольку ключевым моментов в реальных задачах обнаружения аномалий является выбор признаков, которые характеризуют те или иные отклонения от нормы, алгоритмы из ансамбля строят пытаясь угадать хорошие пространства. Здесь популярны:

- Feature Bagging (не очень удачное название) – для каждого алгоритма берут случайное признаковое подпространство,
- Rotated Bagging – в выбранном случайном признаковом подпространстве совершают случайный поворот.

Кстати, здесь «усреднение» не обязательно означает среднее арифметическое всех оценок, интуитивно понятно, что часто может сработать максимум (если какой-то алгоритм уверен в аномальности объекта, то скорее всего так оно и есть).

Приложения поиска аномалий

1. В задачах поиска аномалий важно понимать, как работают алгоритмы поиска. Например, заказчик хочет средство для детектирования поломок, если подойти к решению задачи как к поиску аномалий, то получился алгоритм детектирования «неправильного функционирования оборудования». Он будет «давать сигнал тревоги» не только в случае поломок, но и в случае некорректной эксплуатации прибора, а также при работе в очень редких режимах. Некоторые поломки (очень частые) он может пропускать, т.к. «они уже стали для прибора нормой». Понятно, что при наличии большой размеченной выборки таких проблем не возникает, но на практике оборудование работает не слишком долго, поломок происходит мало (и не все возможные случаются), а некоторые поломки не замечают или замечают с запозданием. Кроме того, некоторые поломки никак не отражаются на показаниях датчиков (и детектировать по показаниям их невозможно).

2. Детекторы аномалий (как и кластеризаторы) можно использовать как генераторы новых признаков. Например, добавлять в матрицу данных оценку аномальности, полученную изолирующим лесом. Заметим, что при этом

- мы не подглядываем в целевые значения, поэтому нет риска переобучения,
- у нового признака естественная интерпретация – насколько объект типичен.

3. В глубоком обучении есть свои методы обнаружения аномалий. В большинстве, они эксплуатируют модельный подход. Например, строится нейросеть, которая прогнозирует временной ряд, точки которые сильно отличаются от прогноза объявляются аномалиями. Или получают компактное представление объектов, для этого используют, например, автокодировщики.

На рис. XX.23 показаны примеры изображений из современного набора данных. Нормальные объекты здесь – фотографии микросхем, капсул, макаронин и т.п. Аномальные – фотографии с повреждёнными объектами: микросхемы с погнутыми ножками, несимметричные капсулы, макаронины со сколами и вкраплениями и т.п. Понятно, что детекторы аномалий, в частности, могут помочь контролировать качество продукции.

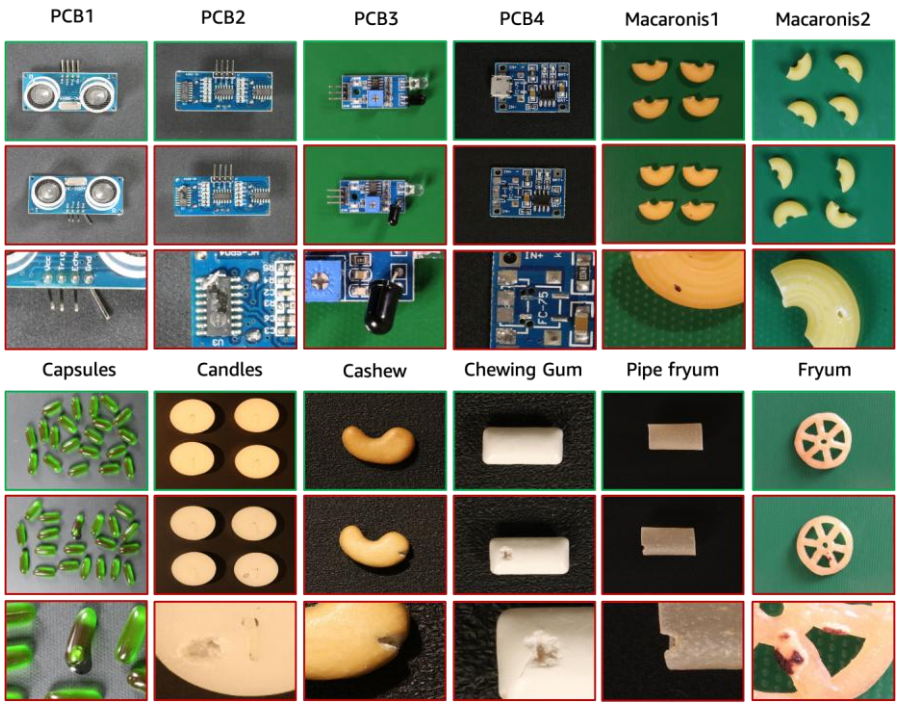

Рис. XX.23. Изображения из датасета VisA (Visual Anomaly Dataset[1]).

4. Отметим также, что бывают и нестандартные аномалии, например коллективные выбросы (Collective outliers[2]). На рис. XX.24 значения пары признаков равномерно покрывают сетку $\{0,\ldots,5\}\times\{0,\ldots,3\}$ с точностью до небольшого шума, при этом паре $(4, 2)$ соответствует существенно больше объектов, чем другим парам. Точки из подобных скоплений и называются коллективными выбросами. В главе «Предобработка данных» рассказывается о нахождении пропущенных значений, в частности, в данных пропуски могут выглядеть как коллективные выбросы (например, если доход клиента неизвестен, то поле заполняется значением 99999).

[1] https://paperswithcode.com/dataset/visa

[2] См. обзор Smiti A. A critical overview of outlier detection methods //Computer Science Review. – 2020. – Т. 38. – С. 100306.

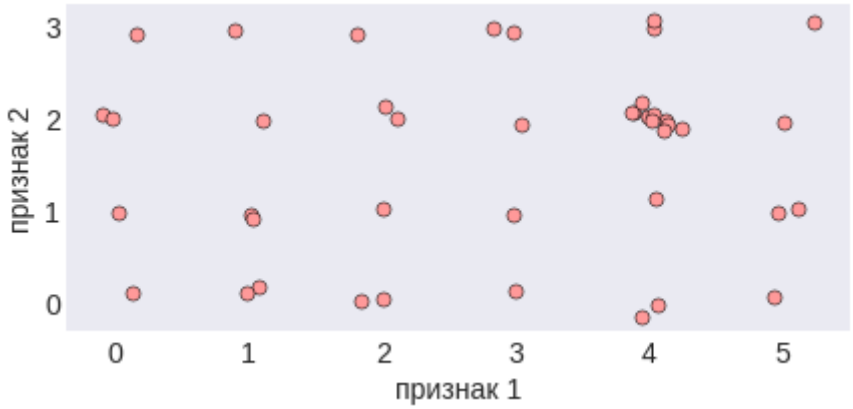

Рис. XX.24. Пример коллективных выбросов.

Вопросы и задачи

1. На иллюстрациях разделяющая линия не соответствует ответам детектора выбросов, предложите способ исправления этого[1].

2. На рис. XX.15 проиллюстрирован подход, в котором оценка аномальности зависит от расстояния до центра кластера. Мы описали идею подхода, но не коснулись важной проблемы выбора числа кластеров. Предложите и реализуйте ансамбль детекторов при разном числе кластеров. Подумайте, как в этом случае лучше агрегировать ответы детекторов.

3. Предложите методы для детектирования коллективных выбросов.

[1] Код можно взять по ссылке: https://github.com/Dyakonov/ml_hacks/blob/master/dj_oneclass_press.ipynb

Поиск аномалий: итоги

Аномалии бывают выбросами (есть в выборке) и новизной (скоро появятся), в слабом смысле (шумами) и сильном – которые обычно и детектируют.

У обнаружения аномалий много применений в задачах, где возможно внезапное нежелательное нештатное поведение системы / показаний. Как правило, в таких задачах нет разметки или есть лишь несколько представителей класса 1 (аномалия), наблюдается большой дисбаланс (аномалий немного).

Основные группы методов: статтесты, моделирование, эвристические итерационные, метрические, сведение к другой задаче (чаще кластеризации и классификации), отдельные алгоритмы ML: метод опорных векторов с одним классом, изолирующий лес (очень хороший и простой метод), эллипсоидальная аппроксимация, а также ансамбли.

Спасибо за внимание к книге!
Замечания по содержанию, замеченные ошибки и неточности можно написать в телеграм-чате
https://t.me/Dyakonovsbook